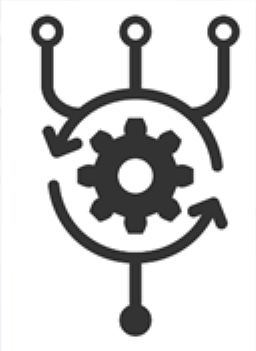
Fairness in Machine Learning - Overview

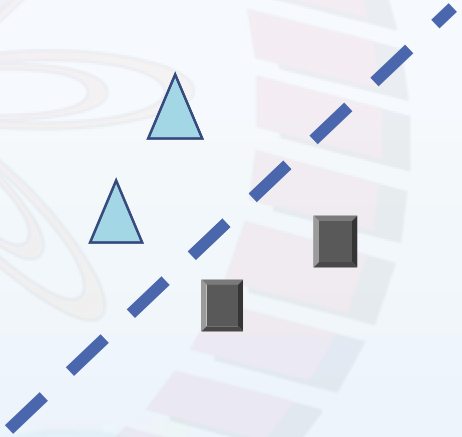
Arun Rajkumar
Dept of DSAI, IITM

Machine Learning Pipeline

Data



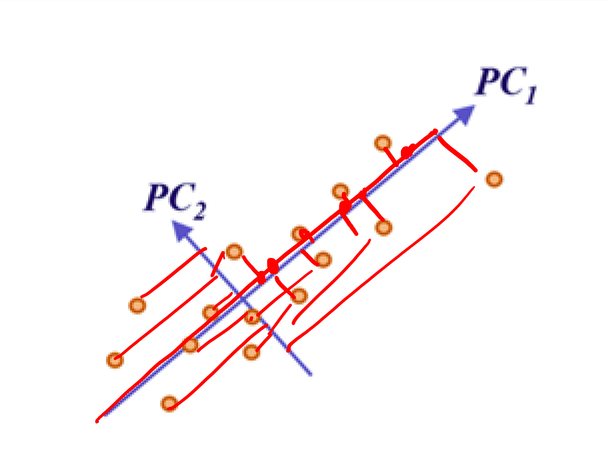
Algorithm



Model


RECAP

Principal Component Analysis - Recap



Input Dataset

28

28

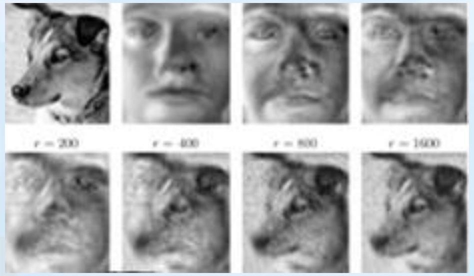
28×28

Principal Components

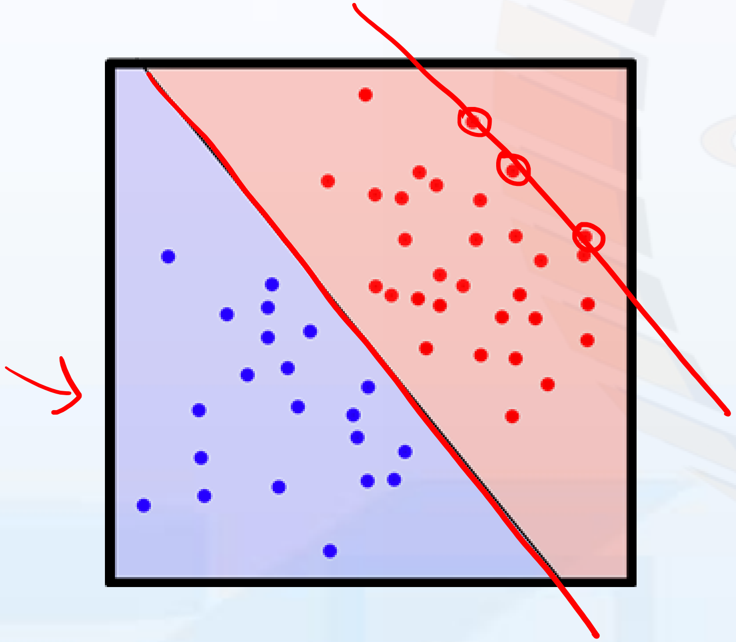
PC1 PC2 PCk.

Test Image

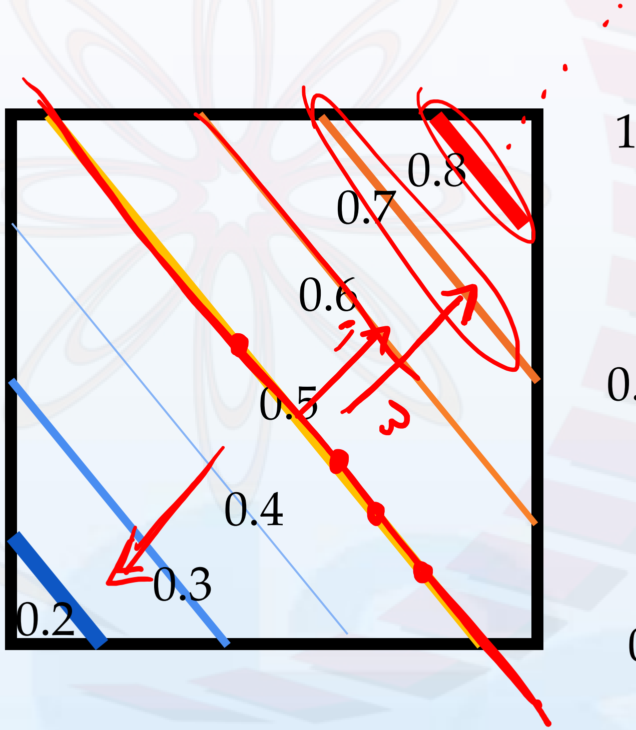
Another Test Image



Logistic Regression - Recap

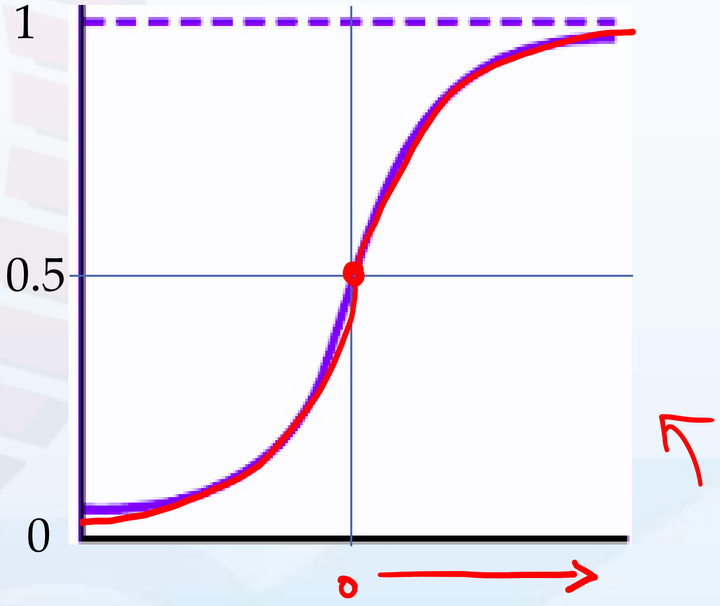


Linearly Separable Dataset



Probabilisitc Model

$P(y=+1 \,(\text{red}) \,|\, x)$



Logistic Function

$P(y=1/x) = 1/(1+\exp(\bar{w}.x))$

Logistic Regression - Recap

$$P(y=+1/x) = g(z)$$

- $g(z) = 0.5$ if $z=0$
- $g(z) \to 1$ as $z \to \infty$
- $g(z) \to 0$ as $z \to -\infty$

Logistic Regression - Recap

Model : LOGISTIC REGRESSION

$$P(y=1/x) = \frac{1}{1+e^{-W^T x}}$$

Data : $\{(x_1, y_1), \ldots, (x_n, y_n)\}$

- How to find W?
- Maximum likelihood!

$$\mathcal{L}(W; \text{Data}) = \prod_{i=1}^{n} \left(g(W^T x_i)\right)^{y_i} \cdot \left(1 - g(W^T x_i)\right)^{(1-y_i)}$$

Logistic Regression - Recap

- No closed form solution for maximisation
- Can perform gradient descent

Gradient

$$\nabla \log L(w) = \sum_{i=1}^{n} x_i \left(y_i - \frac{1}{1+e^{-w^T x_i}} \right)$$


FAIR
YES
NO
?

ugliest language in india

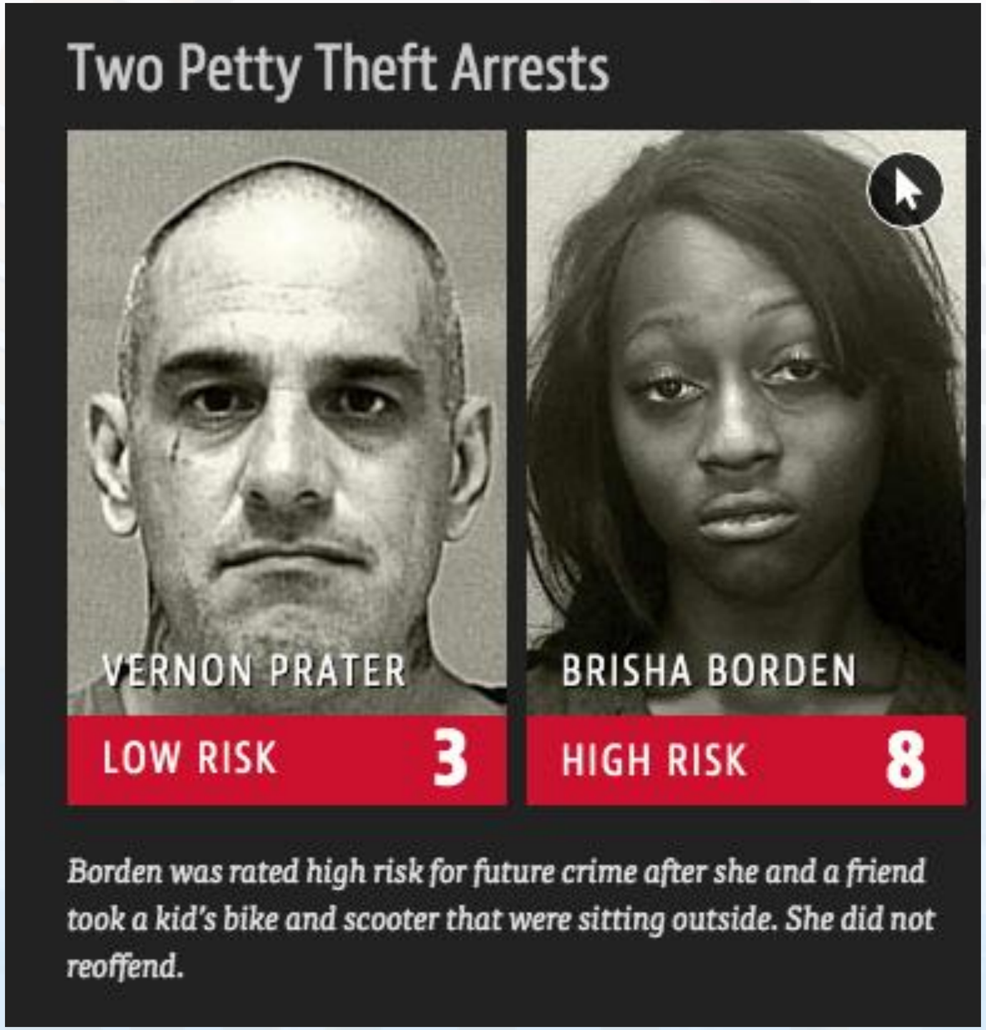
All Videos Images News Shopping

Kannada

What is the **ugliest language in India**? The answer is Kannada, a **language** spoken by around 40 million people in south **India**.

Google apologises for Kannada is the 'ugliest language in India' search result after backlash

"Sometimes, the way content is described on the Internet can yield surprising results to specific queries. We know this is not ideal, but we take swift corrective action when we are made aware of an issue and are continually working to improve our algorithms. Naturally, these are not reflective of the opinions of Google, and we apologize for the misunderstanding and hurting any sentiments," a Google spokesperson said in a response to Hindustan Times.



Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

REUTERS

Amazon scraps secret AI recruiting tool that showed bias against women

Discrimination in Online Ad Delivery

Latanya Sweeney
Harvard University
latanya@fas.harvard.edu

January 28, 2013[1]

Problem Statement

Given online searches of racially identifying names, show that associated personalized ads suggestive of an arrest record do not differ by race.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2208240

Biases in Word Embeddings

Wikpedia

Sexist prejudice			
Profession		Sentiment	
Woman	Man	Woman	Man
Nurse	Officer	Wedding	Reinforcement
Secretary	Hunter	Divorce	Attack
Teacher	Commander	Anulment	Combat
Saleswoman	Guard	Engagement	Power
Actress	Cameraman	Marry	Decrease

Social Media

Sexist prejudice			
Profession		Sentiment	
Woman	Man	Woman	Man
Nurse	Policeman	Agitation	Robber
Secretary	Musician	Mature	Attacker
Pharmacist	Priest	Love	Injured
Religion teacher	Coach	Increase	Fascist
Correspondent	Paramedic	Stubborness	Overwhelmed

https://blog.acolyer.org/2020/12/08/bias-in-word-embeddings/

ChatGPT

Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →

Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests

Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

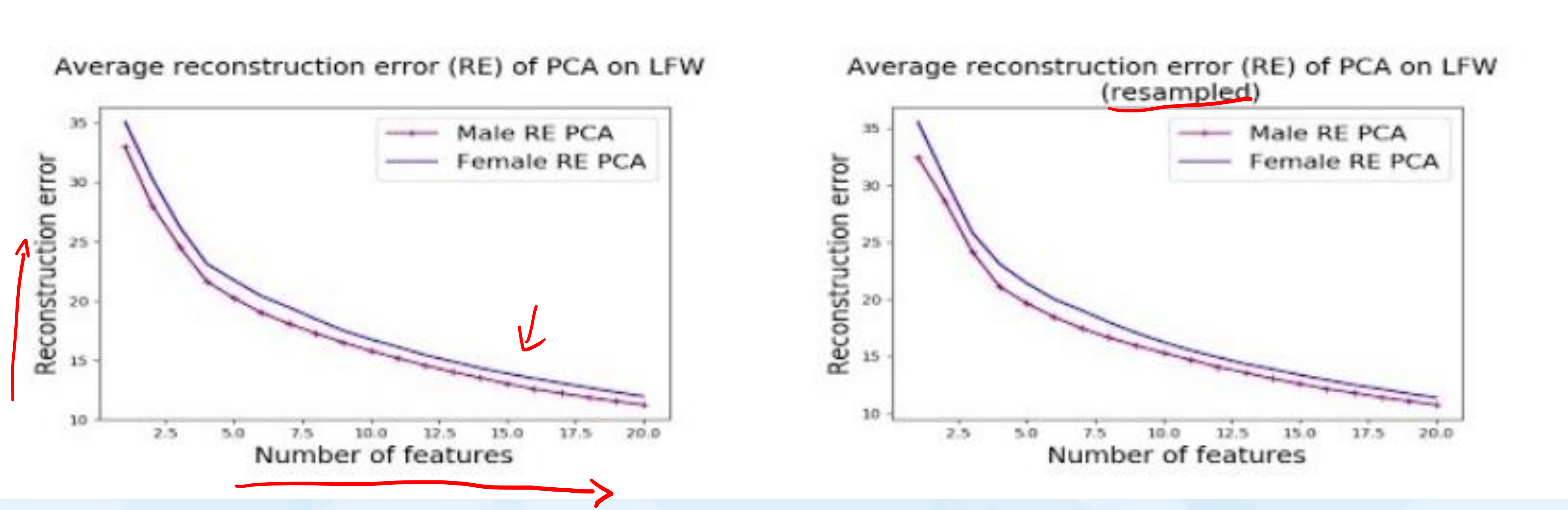
Limited knowledge of world and events after 2021

How do we fix it?

But first, what exactly is fairness?

Fair unsupervised learning

~~Un~~Fair PCA



https://sites.google.com/site/ssamadi/home/fair-pca-homepage

~~Un~~Fair PCA

original data

$$\| X - \hat{X} \|_F^2$$

reconstructed/ approx data.

$$\begin{array}{l} \text{male} \\ \text{female} \end{array} \begin{bmatrix} A \\ B \end{bmatrix} \begin{array}{c} \rightarrow \\ \rightarrow \end{array} \quad U = \begin{bmatrix} U_A \\ U_B \end{bmatrix}$$

Definition: Loss of a population A when approximated by matrix U_A. Let $\hat{A}$ be the optimal rank-d approximation of A.

$$Loss_A(U_A) = Error_A(U_A) - Error_A(\hat{A}) = \|A - U_A\|_F^2 - \|A - \hat{A}\|_F^2$$

https://sites.google.com/site/ssamadi/home/fair-pca-homepage

~~Un~~Fair PCA

$x = \begin{bmatrix} A \\ B \end{bmatrix} \rightarrow U = \begin{bmatrix} U_A \\ U_B \end{bmatrix}$

Fair PCA problem:

$$\min_{U:rank(U)=d} \max \left\{ \frac{1}{|A|} Loss_A(U_A), \frac{1}{|B|} Loss_B(U_B) \right\}$$

https://sites.google.com/site/ssamadi/home/fair-pca-homepage

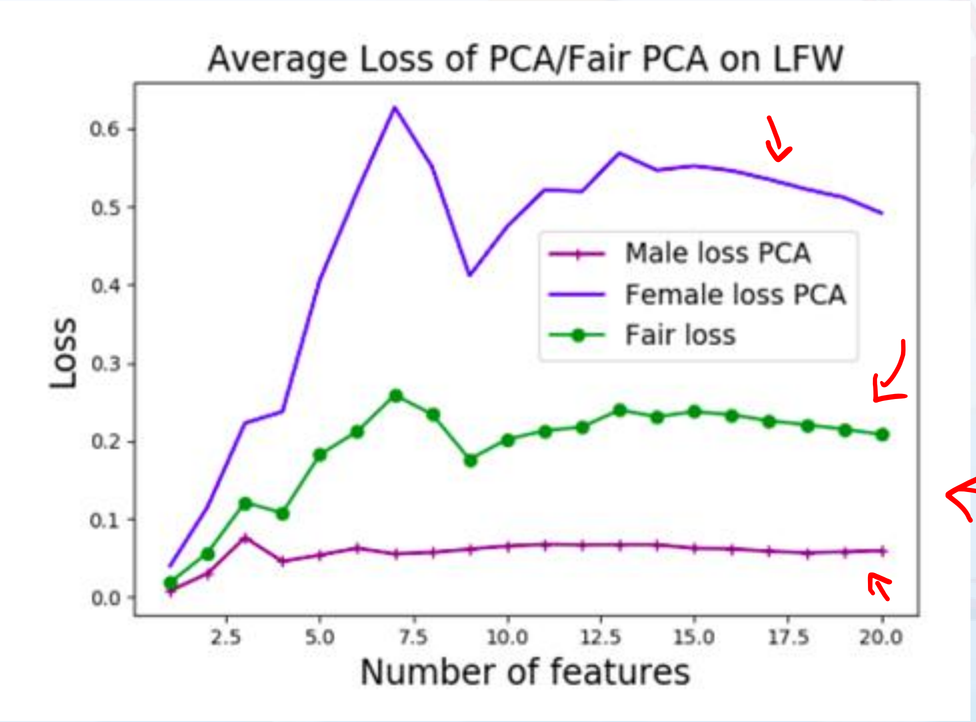
Theorem *Let U be a solution to the Fair PCA problem (1), then*

$$\frac{1}{|A|} loss(A, U_A) = \frac{1}{|B|} loss(B, U_B).$$

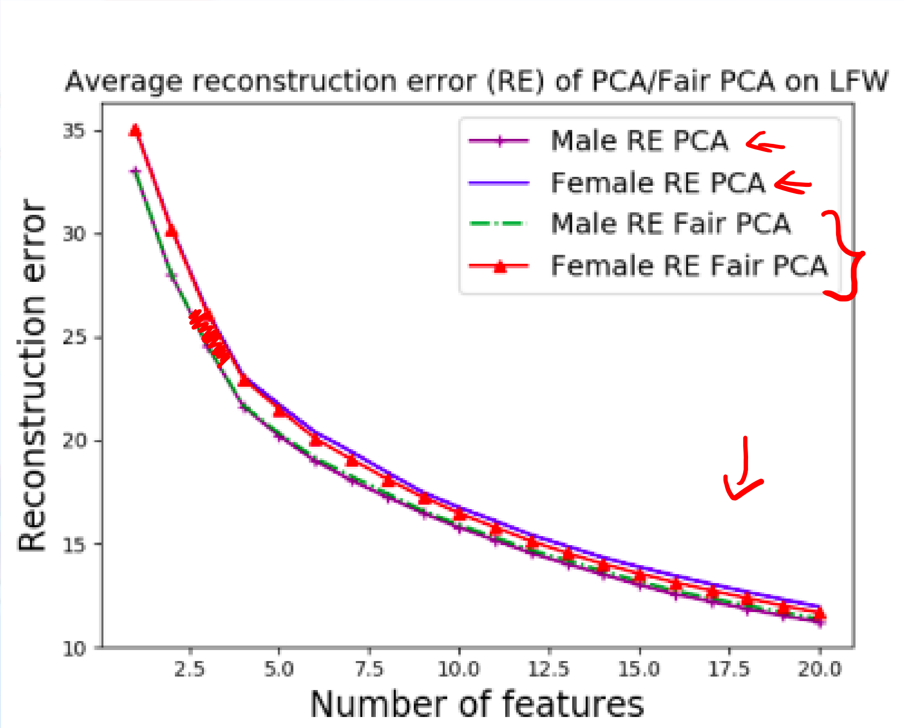
$U = \begin{bmatrix} U_A \\ U_B \end{bmatrix}$

Theorem 5.1. *There is a polynomial-time algorithm that outputs an approximation matrix of the data such that it is either of rank d and is an optimal solution to the fair PCA problem OR it is of rank d + 1, has equal losses for the two populations and achieves the optimal fair PCA objective value for dimension d.*

https://sites.google.com/site/ssamadi/home/fair-pca-homepage



Average Loss of PCA/Fair PCA on LFW
Loss
Number of features
Male loss PCA
Female loss PCA
Fair loss
0.0
0.1
0.2
0.3
0.4
0.5
0.6
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0



Average reconstruction error (RE) of PCA/Fair PCA on LFW
Reconstruction error
Number of features
Male RE PCA
Female RE PCA
Male RE Fair PCA
Female RE Fair PCA
10
15
20
25
30
35
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

Fair Supervised learning

No single answer

Statistical parity
Group fairness
Demographic parity
Conditional statistical parity
Equal opportunity
Equalized odds
Conditional procedure accuracy equality
Disparate mistreatment
Balance for positive class
Balance for negative class
Predictive equality
Conditional use accuracy equality
Predictive parity
Calibration

Statistical Parity

X – set of all data points (people)
C – Set of all data points (people) belonging to a certain protected group
M: **X -> {0,1}** - a classifier, say *logistic regression*

Bias of classifier for the protected group:
Parity(M,C) = Pr(M(x) = 1/x in C) - Pr(M(x) = 1)

Ideal classifier => Parity = 0
In practice, min |parity(M,C)}

Equality of Opportunity

X – set of all data points (people)
C – Set of all data points (people) belonging to a certain protected group
M: **X -> {0,1}** - a classifier, say *logistic regression*.

opportunity_inequality(M,C) = Pr(M(x) = 1 | y=1 and C) - Pr(M(x) = 1 | y=1)

Ideal classifier: opportunity_inequality = 0
In practice: min |opp_ineq(M,C)|

Individual Fairness

X – set of all data points (people)
M **X -> Δ** (simplex) - a classifier, say *logistic regression*.
d: X * X -> R - distance function
D: Δ * Δ -> R – distance between probability vectors

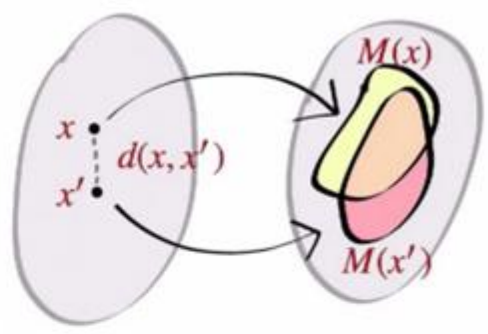


Image source: Moritz Hardt,
Fairness in Machine Learning, NeurIPS 2017

D(M(x),M(x′)) <= d(x,x;′) for any x.x′

Outcome Test (Predictive Parity)

X – set of all data points (people)
C – Set of all data points (people) belonging to a certain protected group
M: **X -> {0,1}** - a classifier, say *logistic regression*.

$$\text{Pred\_parity}(M,C) = \Pr(y=1 \mid M(x) = 1 \text{ and } C) - \Pr(y=1 \mid M(x) = 1)$$

Ideal classifier: Pred_parity = 0
In practice: min |Pred_parity(M,C)|

Impossiblity Result

$$\text{Opportunity\_inequality}(M,C) = \Pr(M(x) = 1 \mid y=1 \text{ and } C) - \Pr(M(x) = 1 \mid y=1)$$

$$\text{Opportunity\_inequality}(M,C) = \Pr(M(x) = 1 \mid y=1 \text{ and } C) - \Pr(M(x) = 1 \mid y=1)$$

Calibration within groups – $\Pr(M(x) = 1 \mid x \text{ in } C) \sim= \Pr(y=1 \mid x \text{ in } C)$

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg [*] Sendhil Mullainathan [†] Manish Raghavan [‡]

Fair Logistic Regression

https://dl.acm.org/doi/fullHtml/10.1145/3308560.3317584

Parity = |Pr(M(x) = 1|C = 1) - Pr(M(x) = 1| C= 0)|

Regularized logistic loss on dataset

$$\underset{w}{\text{minimize}} \quad f_D(\mathbf{w})$$
$$\text{subject to} \quad g_D(\mathbf{w}) \le \tau, \; g_D(\mathbf{w}) \ge -\tau,$$

equivalently.

$$\boxed{|g_D(w)| \le \tau}$$

$$\left\{ w : \left| \sum_{i=1}^{n} (w^T x_i)(s_i - \bar{s}) \right| \le \tau \right\}$$

$s^T e$.

$$g_D(\mathbf{w}) = \sum_{i=1}^{n} (s_i - \bar{s}) \mathbf{x}_i^T \mathbf{w}$$

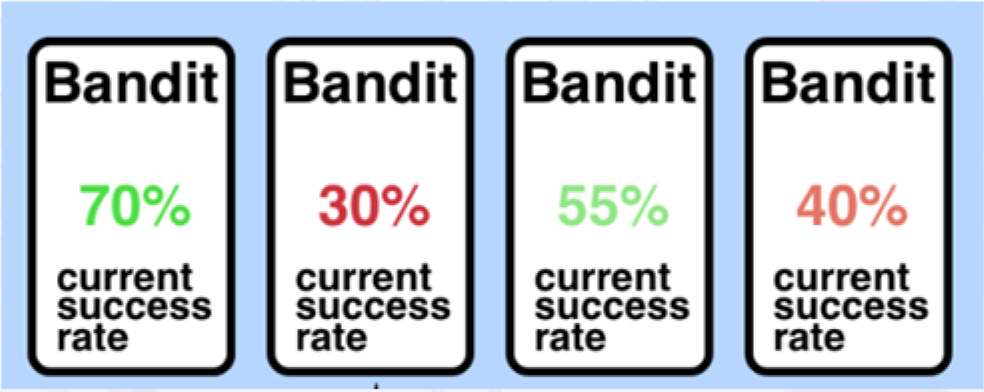
Measures how correlated are the prediction probabilities to the Protected attribute.

Protected attribute for data point i

$$s = \begin{bmatrix} s_1 - \bar{s} \\ s_2 - \bar{s} \\ \vdots \\ s_n - \bar{s} \end{bmatrix} \qquad \begin{bmatrix} w^T x_1 \\ w^T x_2 \\ \vdots \\ w^T x_n \end{bmatrix} = e$$

$s^T e$

Fair Multi Armed Bandits



0.8 0.4 0.5 0.9

Case in point: Swiggy/Zomato wants to assign partners to orders.
Goal: Maximize expected reward over time.

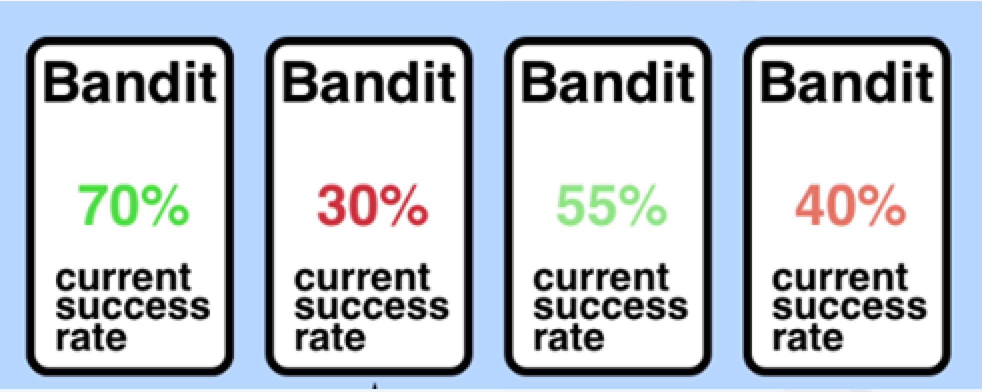
A simple strategy: Pick current best arm with (0.9) probability and uniformly at random with 0.1 prob.

Fair Multi Armed Bandits

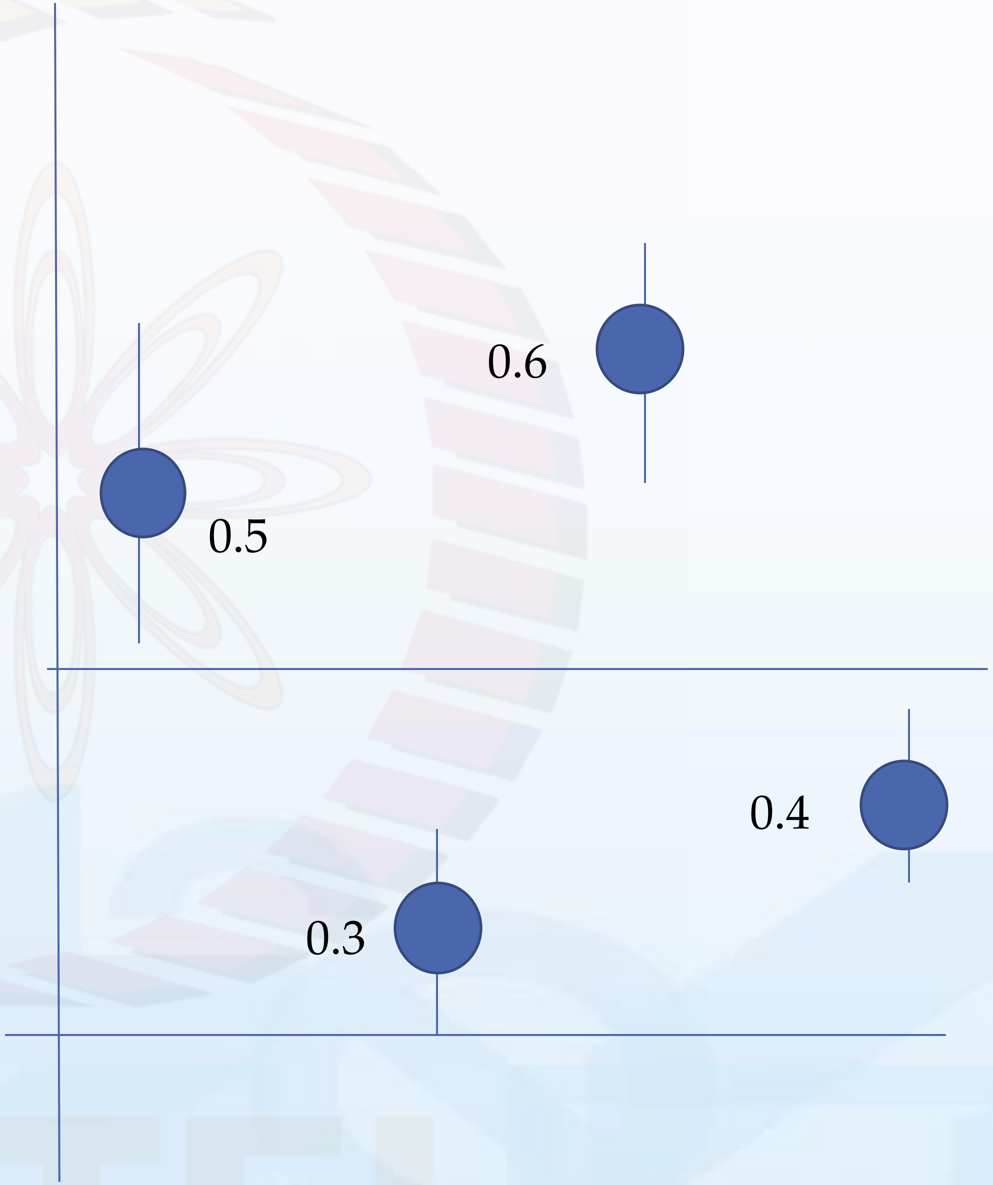
Notion of Fairness:

If partner-a is **truly** better than partner b,
then
Probability that algorithm assigns partners a >=
Probability that algorithm assigns partners b

~~**A simple strategy**: Pick current best arm with (0.9) probability and uniformly at random with 0.1 prob.~~

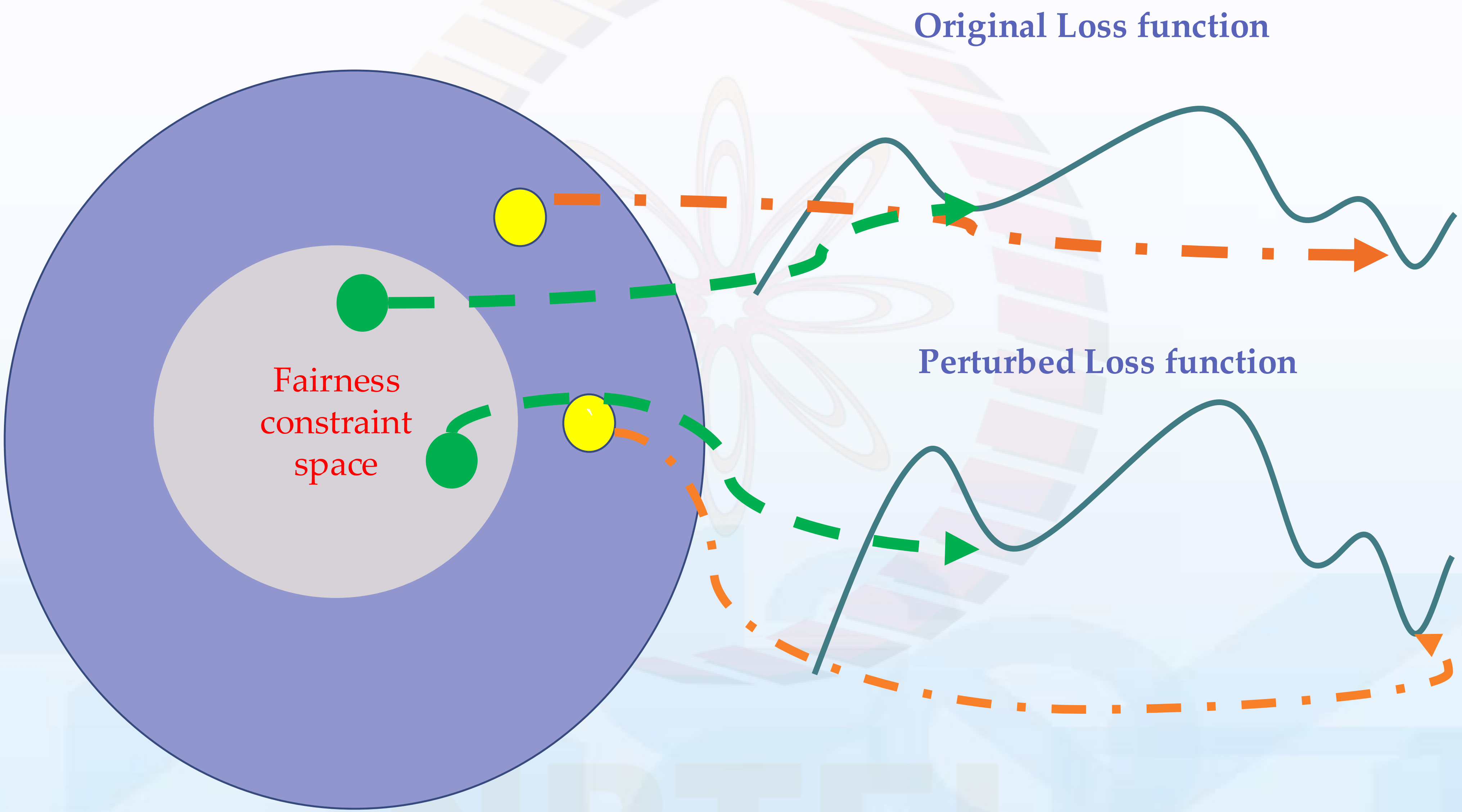


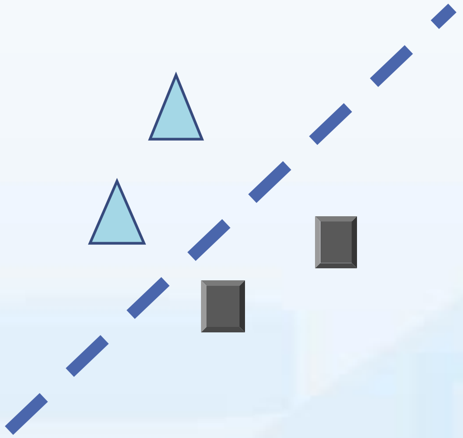
0.5 0.35 0.6 0.38



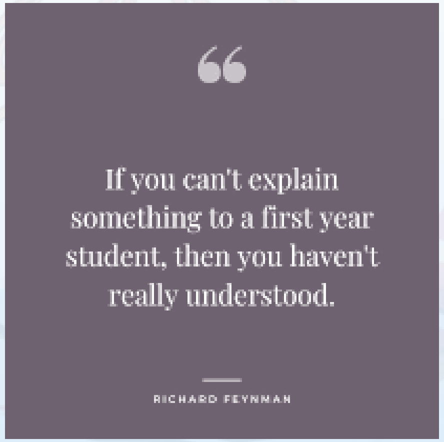
IDEA: Sample uniformly from those whose confidence interval overlaps with the winner

- Fair PCA
- Fair LogReg
- Fair MAB


Original Loss function
Fairness
constraint
space
Perturbed Loss function



Model


If you can't explain something to a first year student, then you haven't really understood.
RICHARD FEYNMAN

Thank you!

Centre for Responsible AI

https://cerai.iitm.ac.in

arunr@cse.iitm.ac.in